INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

A fast algorithm for the two dimensional HJB equation of stochastic control

J. Frédéric BONNANS — Elisabeth OTTENWAELTER — Housnaa ZIDANI

N° 5078

Janvier 2004

THÈME 4

ISSN 0249-6399 ISRN INRIA/RR--5078--FR+ENG

INRIA
ROCQUENCOURT

A fast algorithm for the two dimensional HJB equation of stochastic control

J. Frédéric BONNANS* , Elisabeth OTTENWAELTER† , Housnaa ZIDANI‡

Thème 4 — Simulation et optimisation
de systèmes complexes
Projet SYDOCO

Rapport de recherche n° 5078 — Janvier 2004 — 21 pages

Abstract: This paper analyses the implementation of the generalized finite differences method for the HJB equation of stochastic control, introduced by two of the authors in [4]. The computation of coefficients needs to solve at each point of the grid (and for each control) a linear programming problem.

We show here that, for two dimensional problems, this linear programming problem can be solved in $O(p)$ operations, where p is the size of the stencil. The method is based on a walk on the Stern-Brocot tree, and on the related filling of the set of positive semidefinite matrices of size two.

Key-words: Stochastic control, finite differences, viscosity solutions, consistency, HJB equation, Stern-Brocot tree.

* Projet Sydoco, Inria-Rocquencourt, Domaine de Voluceau, BP 105, 78153 Le Chesnay, France (Frederic.Bonnans@inria.fr).

† IUT de Paris and Projet Sydoco, Inria-Rocquencourt, Domaine de Voluceau, BP 105, 78153 Le Chesnay, France (Elisabeth.Ottenwaelter@inria.fr).

‡ Projet Sydoco, Inria-Rocquencourt and Unité de Mathématiques Appliquées, ENSTA, 32 Boulevard Victor, 75739 Paris Cedex 15, France (zidani@ensta.fr).

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 Le Chesnay Cedex (France)
Téléphone : +33 1 39 63 55 11 —Télécopie : +33 1 39 63 53 30

Un algorithme rapide pour l'équation HJB du contrôle stochastique en dimension deux.

Résumé : Cet article analyse l'implémentation de la méthode de différences finies généralisées pour l'équation HJB du contrôle stochastique, introduite par deux des auteurs dans [4]. Le calcul des coefficients nécessite la résolution en chaque point de la grille (et pour chaque commande) d'un programme linéaire.

Nous montrons que, pour les problèmes bidimensionnels, ce programme linéaire peut se résoudre en $O(p)$ opérations, où p est la taille du stencil. La méthode est basée sur un cheminement dans l'arbre de Stern-Brocot, et sur le remplissage associé de l'ensemble des matrices semidéfinies positives de taille deux.

Mots-clés : Contrôle stochastique, différences finies, solutions de viscosité, consistance, équation HJB, arbre de Stern-Brocot.

AMS Subject classification 93E20, 49L99.

1 Introduction

In this paper we discuss numerical schemes for the HJB equation of stochastic control. The model problem we are considering is

$$(P_{\tau,x}) \qquad \left\{ \begin{array}{l} \mathrm{Min}\, I\!E \displaystyle\int_\tau^T \ell(t,y(t),u(t))\mathrm{d}t + \ell_F(y(T)); \\ \\ \left\{ \begin{array}{l} \mathrm{d}y(t) = f(t,y(t),u(t))\mathrm{d}t + \sigma(t,y(t),u(t))\mathrm{d}w(t), \\ y(\tau) = x, \end{array} \right. \\ \\ u(t) \in U, \;\; \tau \in [0,T], \;\; t \in [\tau,T]. \end{array} \right.$$

Here $T > 0$ is the (given) final time, $y(t) \in I\!R^n$ and $u(t) \in I\!R^m$ are the state and control variable, the latter subject to the constraint $u(t) \in U$ where U is a compact subset of $I\!R^m$ a.e., $\ell : I\!R \times I\!R^n \times I\!R^m \to I\!R$ and $\ell_F : I\!R^n \to I\!R$ are the distributed and final cost, $f : I\!R \times I\!R^n \times I\!R^m \to I\!R^n$ is the trend (deterministic part of dynamics), $\sigma(\cdot,\cdot,\cdot)$ is a mapping from $I\!R \times I\!R^n \times I\!R^m$ into the space of $n \times r$ matrices, and w is a standard r dimensional Brownian motion. The control variable u has to be a function of past events, i.e., is progressively measurable w.r.t. the filtration $\mathcal{F}_t$ associated with the Brownian motion. Let $\mathcal{U}$ be the set of feasible policies, i.e., progressively measurable controls with values in U. We assume for the sake of simplicity that f, σ, ℓ and ℓ_F, are Lipschitz and bounded. Then (e.g. Fleming and Soner [6]) the stochastic differential equation is, for each policy $u \in \mathcal{U}$, well posed and the corresponding expectation $W(t,x,u)$ is well-defined. Denote the transposition operator by $\top$. Let $a(t,x,u) := \frac{1}{2}\sigma(t,x,u)\sigma(t,x,u)^\top$, for all $(t,x,u) \in [0,T] \times I\!R^n \times U$, be the covariance matrix. The value function V of problem $(P_{\tau,x})$, defined by $V(\tau,x) := \inf_u W(\tau,x,u)$, is (P.L. Lions [12]) the unique bounded viscosity solution of the Hamilton-Jacobi-Bellman (HJB) equation

$$\begin{array}{rl} -v_t(t,x) = & \displaystyle\inf_{u\in U} \left\{ \ell(t,x,u) + f(t,x,u)\cdot v_x(t,x) + a(t,x,u) \circ v_{xx}(t,x) \right\}, \\ & \qquad \text{for all } \; t,x \in [0,T] \times I\!R^n. \\ v(T,x) = & \ell_F(x), \;\; \text{for all } \; x \in I\!R^n. \end{array} \qquad (HJB)$$

where v_{xx} denotes the $n \times n$ matrix of second derivatives of v with respect to x, and given two symmetric matrices A, B, of size n, $A \circ B := \sum_{i,j=1}^n A_{ij}B_{ij}$ is the scalar

product associated with the Frobenius norm $\|A\| := (\sum_{i,j=1}^n A_{ij}^2)^{1/2}$ (since we do not use other norms on matrices the notation is non ambiguous). Various numerical methods have been proposed for solving this problem. Classical finite difference methods were discussed in Lions and Mercier [13], see also Menaldi [14]. Markov chain approximation were introduced in Kushner [10], see Kushner and Dupuis [11]. Camilli and Falcone [5] discuss methods based on a priori time discretization (and the related dynamic programming principle for discrete time problems). Krylov [9] gives an error estimate of a large class of discretization schemes. Recent improvements of the error estimates were recently obtained in Barles and Jakobsen [1, 2].

2 Generalized finite differences

Let us recall the generalized finite differences (GFD) method of [4] in the setting of finite horizon problems. The space discretization steps are positive real numbers $h_1, \ldots, h_n$. With a point of the grid $\mathbb{Z}^n$ of coordinate $k \in \mathbb{Z}^n$ is associated the point $x_k := \sum_{i=1}^n k_i e_i$ of the state space, where e_i is the ith standard basis vector. Let $Q \in I\!N$, $Q > 1$ be the number of time steps; set $h_0 := T/Q$ and $t_q := qh_0$, for $q = 0, \ldots, Q$. Denote by v_k^q the approximation of the value function V at $(t, x) = (t_q, x_k)$.

Let $\varphi = \{\varphi_k\}$ be a real valued function over $\mathbb{Z}^n$. The upwind finite difference operator $D^{\pm}$ associated with $f(t_q, x_k, u)$ at point (t_q, x_k) is

$$\left(D^{\pm}\varphi_k\right)_i = \frac{\varphi_{k+e_i} - \varphi_k}{h_i} \quad \text{if} \quad f(t_q, x_k, u)_i \geq 0, \quad \frac{\varphi_k - \varphi_{k-e_i}}{h_i} \quad \text{if not.} \tag{2.1}$$

With $\xi \in \mathbb{Z}^n$, associate the second order finite difference operator

$$\Delta_\xi \varphi_k := \varphi_{k+\xi} + \varphi_{k-\xi} - 2\varphi_k = \varphi_{k+\xi} - \varphi_k - (\varphi_k - \varphi_{k-\xi}). \tag{2.2}$$

The (second-order) stencil $\mathcal{S}$ is a finite set of $\mathbb{Z}^n \setminus \{0\}$ containing $\{e_1, \cdots, e_n\}$. For each $k \in \mathbb{Z}^n$, we perform an approximation of the second-order term in the HJB equation by a linear combination of second order finite difference operators associated with elements of the stencil, i.e., the expression $\sum_{\xi \in \mathcal{S}} \alpha_{q,k,\xi}^u \Delta_\xi v_k^q$ where $\alpha_{q,k,\xi}^u$ are to be set. Let $a^h := \{a_{ij}/h_i h_j\}$ denote the scaled covariance matrix. Following [4] we say that the operator $\sum_{\xi \in \mathcal{S}} \alpha_{q,k,\xi}^u \Delta_\xi$ is a *strongly consistent* approximation of $a(t, x, u) \circ D_{xx}^2$ if

$$\sum_{\xi \in \mathcal{S}} \alpha_{q,k,\xi}^u \xi \xi^\top = a^h(t_q, x_k, u), \quad \text{for all} \quad k \in \mathbb{Z}^n. \tag{2.3}$$

This results in the following explicit (backwards) scheme

$$\begin{aligned} \frac{v_k^q - v_k^{q+1}}{h_0} &= \inf_{u\in U}\left\{ \ell(t_q,x_k,u) + f(t_q,x_k,u)\cdot D^{\pm}v_k^q + \sum_{\xi\in\mathcal{S}} \alpha^u_{q,k,\xi}\Delta_\xi v_k^q \right\} \\ v_k^Q &= \ell_F, \end{aligned} \tag{2.4}$$

for all $q = 0,\ldots,Q-1$ and $k\in\mathbb{Z}^n$. The scheme is monotone (i.e., v_k^q is a non decreasing function of v^{q+1}) if all terms $v_{k'}^q$ in (2.4) appear with nonnegative coefficients. This holds if the coefficients $\alpha^u_{q,k,\xi}$ are nonnegative and, in addition,

$$\sum_{i=1}^n \frac{|f_i(t_q,x_k,u)|}{h_i} + 2\sum_{\xi\in\mathcal{S}} \alpha^u_{q,k,\xi} \le \frac{1}{h_0},\quad \forall\,(k,u)\in\mathbb{Z}^n\times U. \tag{2.5}$$

This last condition ensures the non decrease w.r.t. v_k^q. Since strong consistency implies $\sum_{\xi\in\mathcal{S}}\alpha^u_{q,k,\xi} \le \text{trace}\ \ a^h(t_q,x_k,u)$ by [4, Lemma 2.1], condition (2.5) is satisfied whenever

$$\sum_{i=1}^n \frac{\|f_i\|_\infty}{h_i} + 2\|\,\text{trace}\ \ a^h\|_\infty \le \frac{1}{h_0}. \tag{2.6}$$

Consequently, when $\min_i h_i \downarrow 0$ we may take $h_0 = C\min_i\left(h_i^2\right)^2$, for $C>0$ small enough (depending on f and a), as expected.

If the strong consistency and monotonicity properties holds, then GFD are a particular case of consistent chain Markov approximations, and therefore are convergent in view of Kushner and Dupuis [11, Chapter 10]. Since these schemes are monotone and consistent, convergence of these schemes is also a consequence of Barles and Souganidis [3, Thm 2.1]. It is not difficult to see that this scheme satisfies the hypotheses of Krylov [9], Barles and Jacobsen [1, 2], and hence, the error estimates of these authors apply (for the corresponding adaptation to infinite horizon problems of GDF if necessary).

The interest of GFD is that it eases the analyzis of consistency properties. For instance, [4] provides characterizations of the class of covariance matrices for which the scheme is consistent with the most common stencils, for dimensions $n = 2$ to 4. We say that such matrices are consistent with a given stencil. What remained unclear in the analysis of [4] was the easiness of computing the coefficients $\alpha^u_{q,k,\xi}$. Since coefficients $\alpha^u_{q,k,\xi}$ have to be nonnegative, solving (2.3) amounts to solve linear inequality constraints (equivalently, a linear program with zero cost) which may be

expensive if the stencil is large. Remember that this has to be done at each point of the spatial grid, for each time step (and each control is covariances depend on the control). Define the size of a stencil $\mathcal{S}$ as

$$\text{size}(\mathcal{S}) := \max\{\|\xi\|_\infty;\ \ \xi \in \mathcal{S}\}.$$

The main result of this paper is, for two dimensional problems, an algorithm for computing the coefficients in $O(\text{size}(\mathcal{S}))$ operations. More generally, for nonconsistent problems the algorithm computes the closest consistent matrix (in the Frobenius norm) in $O(\text{size}(\mathcal{S}))$ operations. In addition, it has a recursive property: the closest consistent matrix for stencil of size p is computed in $O(1)$ operations after having obtained the closest consistent matrix for stencil of size $p-1$.

The main result is strongly related to geometric properties of the set of PSD (symmetric, positive semidefinite) matrices on $\mathbb{R}^2$, that are the subject of the next section.

3 Structure of 2D covariance matrices

Scaled covariance matrices belong to the cone $\mathcal{C}$ of PSD matrices. We may view these matrices as elements of $\mathbb{R}^3$. The mapping

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix} \to (a_{11},\ \sqrt{2}a_{12},\ a_{22})^\top \tag{3.7}$$

is norm preserving from the space of 2×2 symmetric matrices, endowed with the Frobenius norm, onto the three dimensional Euclidean space. The image of the PSD cone by the mapping (3.7) is the set

$$\left\{z \in \mathbb{R}^3; z_1 \geq 0;\ z_3 \geq 0;\ \tfrac{1}{2}(z_2)^2 \leq z_1 z_3\right\}. \tag{3.8}$$

It is convenient to represent directions of this cone by drawing their intersection with the hyperplane $z_1 + z_3 = 1$ (image of the set of matrices with unit trace), see figure 1. By the orthonormal change of coordinates

$$w_1 = (z_1 - z_3)/\sqrt{2};\ \ w_2 = z_2;\ \ w_3 = (z_1 + z_3)/\sqrt{2},$$

we obtain that this intersection is the Euclidean ball of $\mathbb{R}^2$ of radius $1/\sqrt{2}$. For a given PSD matrix a, coordinates in this hyperplane are $(w_1, w_2)/w_3 = (a_{11} + a_{22})^{-1}(a_{11} - a_{22}, 2a_{12})$ and are called the *view* of a. The view of matrices with unit

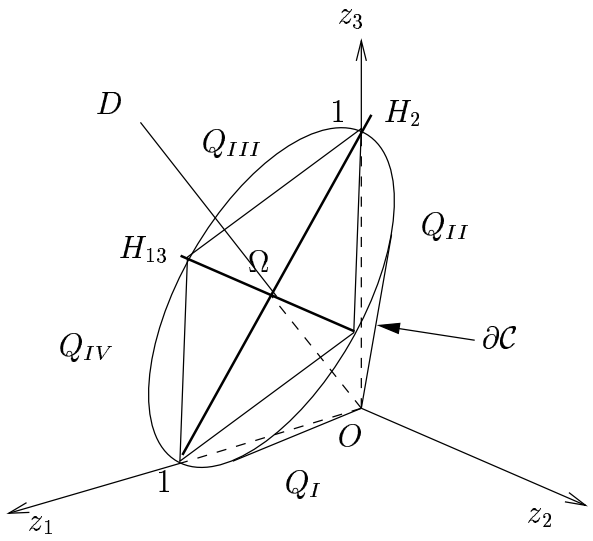

Figure 1: Cone of positive semidefinite matrices

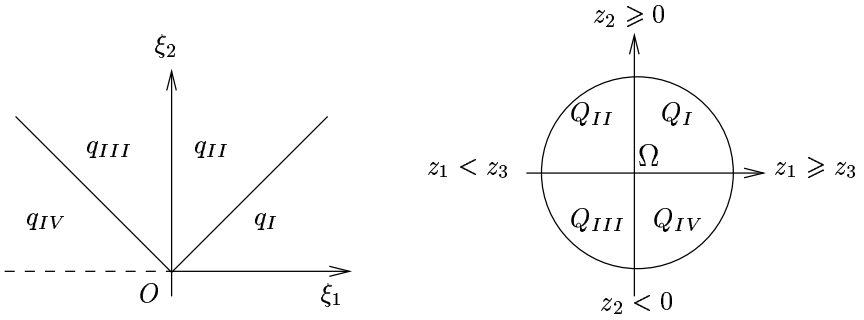

Figure 2: quadrant

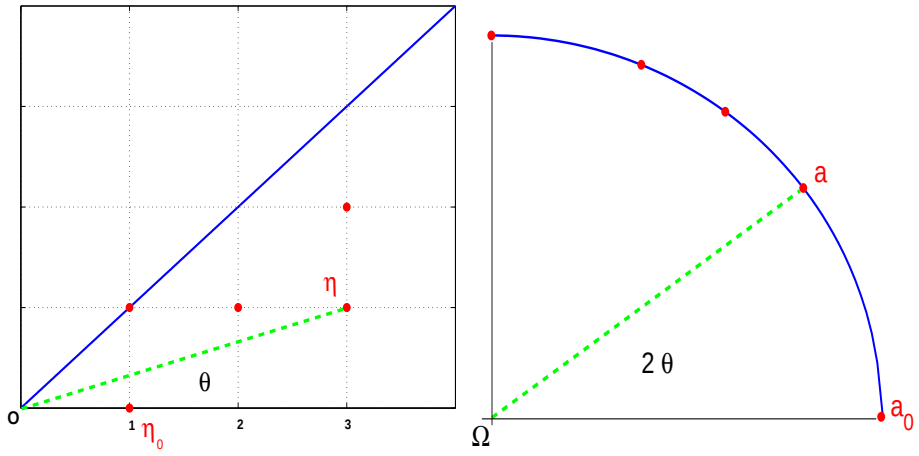

Figure 3: Correspondance of angles

trace is simply $(a_{11} - a_{22}, 2a_{12})$ and the corresponding set is the unit Euclidean ball. The view of the identity, denoted as Ω, is the zero vector, and the view of $\eta\eta^\top$, where $\eta := (1\ 0)^\top$, is $(1\ 0)$.

The lemma below eases the computation of the view of any rank one symmetric nonnegative matrix, and is illustrated in figures 2, 3.

Lemma 3.1 *Let $\eta = (\cos\theta, \sin\theta)$. Then the view of $\eta\eta^\top$ makes an angle of 2θ with the view of $(1,0)(1,0)^\top$.*

Proof. With η are associated $z = (\cos^2\theta, \sqrt{2}\cos\theta\sin\theta, \sin^2\theta)^\top$ and $w = (\cos^2\theta - \sin^2\theta, 2\cos\theta\sin\theta, 1)/\sqrt{2} = (\cos 2\theta, \sin 2\theta, 1)/\sqrt{2}$. The result follows. ■

Let us discuss the case of diagonal dominant matrices. The view of such matrices is the unit ball of $L^1(I\!R^2)$, since it can be easily checked that a matrix is diagonal dominant iff $|a_{11} - a_{22}| + 2|a_{12}| \leq a_{11} + a_{22}$. For a diagonal dominant matrix we have the well-known decomposition

$$\begin{aligned} a = \ & (a_{11} - |a_{12}|) \begin{pmatrix} 1 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \end{pmatrix} + (a_{22} - |a_{12}|) \begin{pmatrix} 0 \\ 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \end{pmatrix} \\ & + \max(a_{12}, 0) \begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \end{pmatrix} + \max(-a_{12}, 0) \begin{pmatrix} -1 \\ 1 \end{pmatrix} \begin{pmatrix} -1 & 1 \end{pmatrix} \end{aligned} \tag{3.9}$$

Let us call "inner region" of the PSD cone, the set of diagonal dominant matrices. There are four outer regions corresponding to the violation of one of the four con-

straints $\pm a_{12} \leq a_{ii}$, for $i = 1, 2$. They are numbered from I to IV according to figure 2. The outer region I is the set of PSD and non diagonal dominant matrices such that $a_{22} < a_{12} < a_{11}$. It is easy to reduce to this case by permutation of variables and change of sign of one state variable. Therefore in the sequel we will discuss essentially the fast decomposition of such matrices. Note that for PSD and diagonal dominant matrices in region I an alternative decomposition, involving the identity matrix, and referred to in section 5, is

$$a = (a_{11} - a_{22}) \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + (a_{22} - a_{12}) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + a_{12} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}. \tag{3.10}$$

4 The Stern-Brocot tree

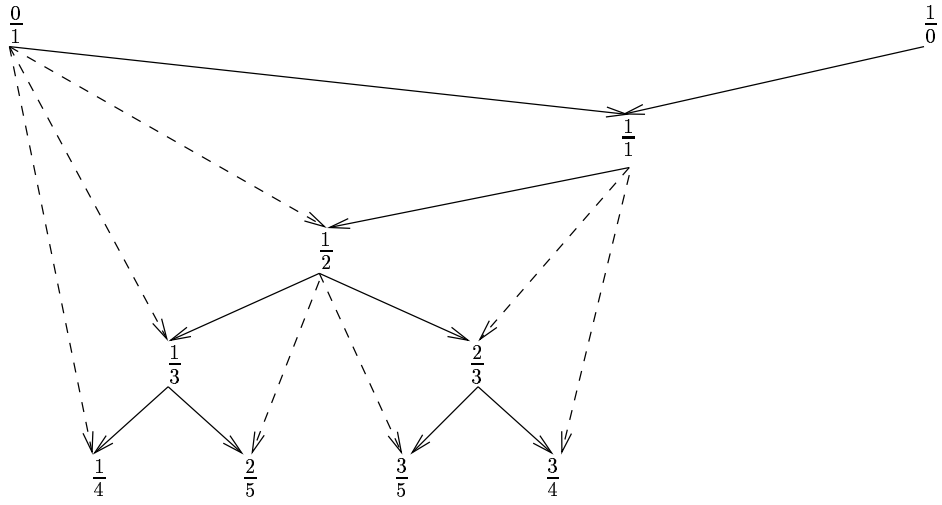

Figure 4: Stern-Brocot tree 1

If the function φ of section 2, defined over $\mathbb{Z}^n$, is the value at grid points of a smooth function $\Phi : I\!R^n \to I\!R$, i.e., $\varphi_k = \Phi(x_k)$, where $x_k := \sum_i k_i h_i$, then the operator Δ_ξ defined in (2.2) allows, as can be seen by a Taylor expansion of Φ around x_k, to obtain a consistent approximation of $\Phi''(x_k)(x_\xi, x_\xi)$, the curvature of Φ at x_k along direction x_ξ. The consistency condition (2.3) expresses the fact that a non-negative combination of such curvatures equals the second order term of the HJB equation. Two elements of the stencil generate the same direction if they are not

linearly independant. Since the algorithm should use points in the stencil as close to x_k as possible, it suffices to take such ξ with relatively prime components.
For two dimensional problems on which we focus now, such points have a specific structure. Assume for simplicity that $k = 0$. For reason of symmetries, we have represented in figure 5 one eighth of the neighbouring points, namely the points ξ in $\mathbb{Z}^2_+$, such that $\xi_2 \leq \xi_1$. Those with an irreducible associated (symbolic) fraction ξ_2/ξ_1, that we will call irreducible points, are in red (boldface in black and white).

As we will see, a very effective way for generating these irreducible points is to use the *Stern-Brocot tree* (see [7]) (which by the way is not a tree in the classical sense), displayed in figure 4. In the sequel, when we write q/p this should be understood as the pair (p, q), so that $p = 0$ makes no problem.

The tree starts with two roots $0/1$ and $1/0$. At any stage of the construction, between two adjacent nodes q/p and q'/p', called the parents, insert the son node $(q + q')/(p + p')$. The two roots are adjacent, and hence, the first son is $1/1$. Then each son is made adjacent with each of his two parents, and we can repeat the process of generating sons (in any order).

Figure 5 shows the links between parents and son. For convenience we give a short proof of some classical properties of the Stern-Brocot tree.

Lemma 4.1 *Let q/p and q'/p' be adjacent nodes such that $q/p < q'/p'$, with son q''/p'', where $p'' = p + p'$, $q'' = q + q'$. Then*
(i) $q/p < q''/p'' < q'/p'$,
(ii) *every node of the Brocot tree is irreducible,*
(iii) *every irreducible fraction b/a belongs to the Brocot tree.*
Furthermore, if q/p and q'/p' are adjacent nodes of the tree such that $q/p < b/a < q'/p'$, then

$$a \geq p + p'; \quad b \geq q + q'. \tag{4.11}$$

Proof. (i) It is easily checked that $q/p < (q + q')/(p + p') < q'/p'$. (This property explains why generation of sons may be made in any order.)
(ii) We prove by induction that, if q/p and q'/p' are adjacent nodes of the tree, then

$$q'p - qp' = 1. \tag{4.12}$$

The relation is obviously true for the root nodes $0/1$ and $1/0$. Assume that it is satisfied for adjacent nodes q/p and q'/p'. It follows from (4.12) that $q'(p + p') - p'(q + q') = 1$ and $p(q + q') - q(p + p') = 1$, proving the induction. Combining (4.12)

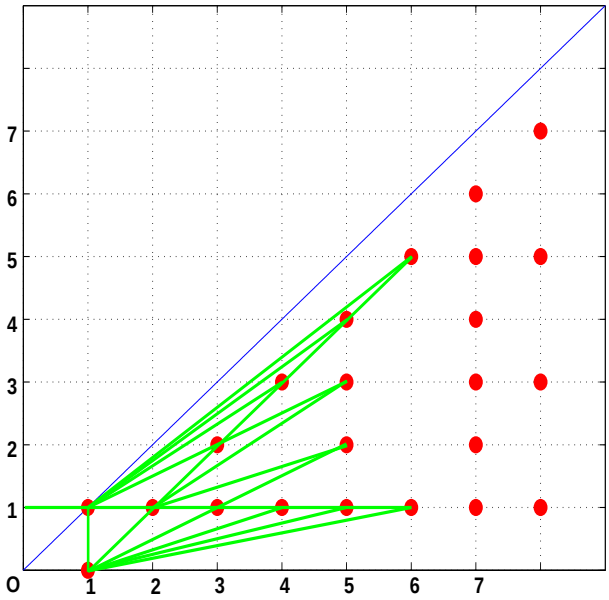

Figure 5: Family relations in regular grid

and Bézout's theorem, we obtain (ii).
(iii) Let b/a be an irreducible fraction, with $0 < b/a < 1$, and q/p, q'/p' be adjacent nodes of the tree such that $q/p < b/a < q'/p'$. Then $bp - aq \geq 1$ and $aq' - bp' \geq 1$. Multiply the first (second) inequality by p' (by p) and add them; multiply the first (second) inequality by q' (by q) and add them; using (4.12), relation (4.11) follows. Since $p'' \geq \max(p, p') + 1$, this relation implies that there is a finite number of couple of adjacent nodes $(q/p, q'/p')$ in the tree such that $q/p < b/a < q'/p'$ holds. This is the case for the two root nodes. Assume now that b/a does not belong to the Stern-Brocot tree. If $q/p < b/a < q'/p'$, setting $q'' = q + q'$ and $p'' = p + p'$, we see that either $q/p < b/a < q''/p''$, or $q''/p'' < b/a < q'/p'$. In this way we generate an infinite sequence of adjacent nodes such that $q/p < b/a < q'/p'$. The desired contradiction follows. ■

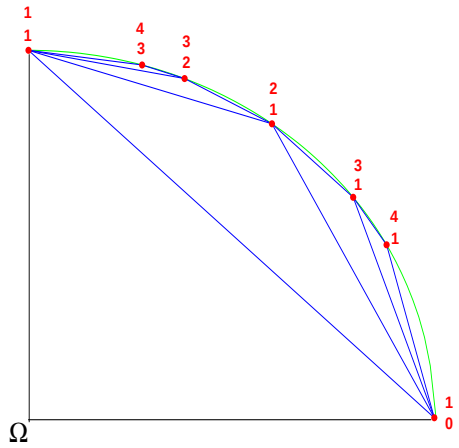

Figure 6: Correspondance of directions

5 Decomposition of the scaled covariance matrix

As discussed at the end of section 3, it suffices to discuss the case when the matrix a^h is in the outer region I; i.e., when it is PSD and non diagonal dominant, and $a_{22} < a_{12} < a_{11}$. On figure 6, this means that the view of a^h belongs to the quarter of ball in the upper right side, and is not in the triangle with summits of coordinates $(0,0)$, $(1,0)$ and $(0,1)$, corresponding to the identity matrix, and degenerate diffusions with horizontal and angle of $\pi/4$ diffusions. (The cone generated by these three points is the set of matrices for which decomposition (3.10) holds).

With every node q/p of the Stern-Brocot tree, $q \leq p$, we associate the directions $\xi_{p,q} := (p \;\; q)^\top$ and $X_{p,q} := \xi_{p,q}\xi_{p,q}^\top$. With two adjacent nodes is associated the plane $H(q/p, q'/p')$ generated by $X_{p,q}$ and $X_{p',q'}$, and two half spaces, the inner one (containing the identity matrix) and the outer one. Denote by $P_H(q/p, q'/p')$ the orthogonal projection onto this plane (since the mapping onto $\mathbb{R}^3$ is norm invariant, projection w.r.t. Frobenius norm is equivalent to the Euclidean projection in the image space $\mathbb{R}^3$).

Beginning the search of a decomposition, we are in the following situation: the matrix a^h belongs to the outer half space of $H(q/p, q'/p')$, with $q/p = 0/1$ and $q'/p' = 1/1$. So, let us assume more generally that a^h belongs to the outer half space of $H(q/p, q'/p')$, where q/p and q'/p' are adjacent nodes. Note that its projection

on the cone generated by matrices of the form X_{p_i,q_i}, with either $q_i/p_i < q/p$ or $q'/p' < q_i/p_i$, belongs to the cone generated by $X_{p,q}$ and $X_{p',q'}$.

In that case we should use another direction of the form $\hat{q}/\hat{p}$, with $\hat{q}$ and $\hat{p}$ nonnegative, such that $q/p < \hat{q}/\hat{p} < q'/p'$, and as small as possible. In view of (4.11), the optimal choice is to take the son $q''/p'' = (q+q')/(p+p')$ (i.e. $q''/p'' = 1/2$ the first time). Then (see figure 6) there are two possibilities.

- The matrix a^h belongs to both inner half spaces of $H(q/p, q''/p'')$ and $H(q''/p'', q'/p')$. Then a^h belongs to the cone generated by $X_{p,q}$, $X_{p',q'}$ and $X_{p'',q''}$. Since these three matrices are linearly independant, the corresponding coefficients are solution of the invertible (three dimensional) system

$$\alpha_{p,q}\ X_{p,q} + \alpha_{p',q'}\ X_{p',q'} + \alpha_{p'',q''}\ X_{p'',q''} = a^h. \tag{5.13}$$

- The matrix a^h belongs to at least one outer half space. Since $X_{p'',q''}$ belongs to the boundary of the cone of PSD matrices, a^h cannot belong to both outer half spaces (see figure 6). We are therefore lead to the situation at the beginning, setting either q/p or q'/p' to q''/p''.

This leads to an effective algorithm, that will stop either if the exact decomposition is obtained, or if either $p'' > p_{max}$, or if the projection of a^h onto $H(q/p, q'/p')$ is close enough to a^h. The precise algorithm is as follows; ε is the maximal relative error of projection of a^h onto the class of consistent matrices, and p_{max} is the size of stencil:

Algorithm DECOMP

INITIAL PHASE: Data $\varepsilon > 0$, p_{max}. Set $k := 0$.

- If a^h is diagonal dominant: set α using (3.10) and stop.
- Reduction to region I, i.e. $a^h_{22} < a^h_{12} < a^h_{11}$.
  Set $q_0/p_0 := 0/1$, $q'_0/p'_0 := 1/1$.

REPEAT

- Compute $a' := P_H(q/p, q'/p')a^h$.
- If $\|a' - a^h\| \leq \varepsilon \|a^h\|$ or $p + p' \geqslant p_{max}$: compute α, decomposition of a' as combination of $X_{p,q}$ and $X_{p',q'}$ and stop.
- Set $q''/p'' := (q+q')/(p+p')$.
- If a^h in inner half spaces of $H(q/p, q''/p'')$ and $H(q/p, q''/p'')$: compute α using (5.13) and stop.
- If a is in outer half space of $H(q/p, q''/p'')$: $q'/p' := q''/p''$.
  Otherwise $q/p := q''/p''$.

- $k := k + 1$.

END REPEAT

From the above discussion we have the following result.

Theorem 5.1 *Algorithm* **DECOMP** *stops after no more than p_{max} iterations, the cost at each iteration is $O(1)$ operations, and hence, its total cost is no more than $O(p_{max})$.*

Obviously it is useful to compute the largest distance between a^h and its projection (as a function of p_{max}) and to evaluate the resulting approximation error. This is the subject of the next section.

6 Projection errors for scaled covariance matrices

Let $\mathcal{S}_p$ denote the stencil of size p reduced to irreducible elements:

$$\mathcal{S}_p := \{(\xi_1, \xi_2) \in \mathbb{Z} \times I\!N;\ \max(|\xi_1|, \xi_2) \leq p;\ (|\xi_1|, \xi_2) \text{ irreducible}\}\ .$$

(the point $(0,0)$ is considered as not irreducible here). The polyhedral cone generated by these directions is $\mathcal{C}(\mathcal{S}_p) = \{\sum_{\xi \in \mathcal{S}_p} \alpha_\xi \xi\xi^\top;\ \alpha_\xi \geq 0\}$. By $\lceil r \rceil$ we denote the smallest integer greater than r.

Lemma 6.1 *The distance from a PSD matrix a to $\mathcal{C}(\mathcal{S}_p)$ is at most $\varepsilon_p \|a\|$, where*

$$\varepsilon_p := \frac{\sqrt{p^2+1} - p}{\sqrt{2}\ \sqrt{2\ p^2+1}} \leq \frac{1}{4}\ p^{-2}. \tag{6.14}$$

Conversely, given $\varepsilon > 0$, the distance from a to $\mathcal{C}(\mathcal{S}_p)$ is at most ε when $p \geq p_\varepsilon$, with

$$p_\varepsilon := \left\lceil \frac{\sqrt{1-\varepsilon^2} - \varepsilon}{2\sqrt{\varepsilon\sqrt{1-\varepsilon^2}}} \right\rceil \tag{6.15}$$

Proof. We may assume that $\|a\| = 1$. Let a' be the projection of a onto $\mathcal{C}(\mathcal{S}_p)$. Let us prove first that, if a' is the projection on the hyperplane spanned by $\xi\xi^\top$ and $\xi'(\xi')^\top$, then

$$\|a - a'\| \quad \leq \quad \frac{\left(1 - \cos(\widehat{\xi, \xi'})\right)}{\sqrt{2} \cdot \sqrt{1 + \cos^2(\widehat{\xi, \xi'})}}\ \|a\| \tag{6.16}$$

the bound being sharp. Indeed, we may assume that $\xi = (\cos\theta \ \sin\theta)^\top$ and $\xi' = (\cos\theta' \ \sin\theta')^\top$. Set $\theta'' := \frac{1}{2}(\theta + \theta')$ and $\xi'' = (\cos\theta'' \ \sin\theta'')^\top$. By reasons of symmetry, the maximal error is reached for $a = \xi''(\xi'')^\top$, with $\xi = (\cos\theta'' \ \sin\theta'')$, and its projection is of the form $a' = \alpha b$, where $b := (\xi\xi^\top + \xi'(\xi')^\top)$, for some $\alpha \in I\!R_+$.

The minimum w.r.t. α of $\|a - \alpha b\|^2$ is

$$\Delta = \|a\|^2 - (a \circ b)^2/\|b\|^2 = 1 - (a \circ b)^2/\|b\|^2.$$

Since this amount is invariant w.r.t. a translation of angles we may assume that $\theta + \theta' = 2\theta'' = 0$, and hence $\theta' = -\theta$, $a = (1, \ 0, \ 0)^\top$, $b = \left(2\cos^2\theta, \ 0, \ 2\sin^2\theta\right)^\top$. We obtain $\|b\|^2 = 4(\cos^4\theta + \sin^4\theta)$ and $a \circ b = 2\cos^2\theta$. It follows that $\Delta = 1 - (a \circ b)^2/\|b\|^2 = \sin^4\theta/(\cos^4\theta + \sin^4\theta)$. Setting $\delta = |\theta' - \theta| = 2|\theta|$, and combining with

$$\begin{aligned} 2\sin^2\theta = &\quad 1 - \cos^2\theta + \sin^2\theta = 1 - \cos\delta \\ \cos^4\theta + \sin^4\theta = &\quad (\cos^2\theta - \sin^2\theta)^2 + 2\cos^2\theta\sin^2\theta = \cos^2\delta + \tfrac{1}{2}\sin^2\delta \end{aligned}$$

we get (6.16).

In the p-stencil, the greatest angle between two consecutive vectors is the angle between $\xi_0 = (1 \ 0)^\top$ and $\xi_1 = (p \ 1)^\top$. By (6.16) and $\cos(\xi_0, \xi_1) = p/\sqrt{p^2+1}$ we have that, in the p-stencil, the largest error is (6.14).

We now prove (6.15). By (6.14), the relative error will be at most ε if $\frac{\sqrt{p^2+1}-p}{\sqrt{2}\ \sqrt{2\ p^2+1}} \leq \varepsilon$. Taking squares in this inequality, we obtain the equivalent relation (since $\varepsilon > 0$)

$$\left(p^2 + 1 + p^2 - 2p\sqrt{p^2+1}\right) \leq 2\varepsilon^2\left(2\ p^2 + 1\right),$$

or $\left(2\ p^2+1\right)\left(1 - 2\ \varepsilon^2\right) \leq 2p\sqrt{p^2+1}$. This inequality having positive sides we again have an equivalent relation by taking squares; the resulting inequality $\left(2\ p^2+1\right)^2\left(1 - 2\ \varepsilon^2\right)^2 \leq 4p^2\left(p^2+1\right)$ reduces to

$$p^4 + p^2 - \frac{\left(1 - 2\ \varepsilon^2\right)^2}{16\ \varepsilon^2\left(1 - \varepsilon^2\right)} \geq 0.$$

This quadratic inequality w.r.t. $q := p^2$ has discriminant

$$\Delta = 1 + \frac{\left(1 - 2\ \varepsilon^2\right)^2}{4\ \varepsilon^2\left(1 - \varepsilon^2\right)} = \frac{1}{4\ \varepsilon^2\left(1 - \varepsilon^2\right)}.$$

The positive root is $q_1 = \dfrac{1 - 2\varepsilon\sqrt{1-\varepsilon^2}}{4\ \varepsilon\sqrt{1-\varepsilon^2}} = \dfrac{(\sqrt{1-\varepsilon^2} - \varepsilon)^2}{4\ \varepsilon\sqrt{1-\varepsilon^2}}$. Therefore (6.14) holds iff $p \geq \sqrt{q_1}$. The result follows. ■

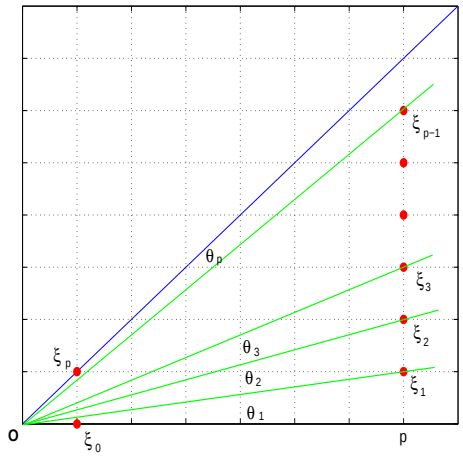

Figure 7: View of maximal error

We display in the table below the first values of ε_p and some values of p_ε. An algorithm involving only the closest neighbour can make up to 17 % of relative error on covariances, and hence, will perform poorly in general. A relative precision of 1 % needs to take $p = 5$. This motivates our effort to make a theory for arbitrary large values of p.

p	ε_p
1	0.169102
2	0.055642
3	0.026325
4	0.015153
5	0.009804
15	0.001109

ε	p
10^{-1}	2
10^{-2}	5
10^{-3}	16
10^{-4}	20
10^{-5}	159
10^{-7}	1 582

Remark 6.1 If consistency does not hold, then the numerical scheme can be interpreted as as consistent approximation for the perturbed HJB equation

$$\begin{aligned} -v_t(t,x) = & \inf_{u\in U}\left\{\ell(t,x,u) + f(t,x,u)\cdot v_x(t,x) + a_p(t,x,u)\circ v_{xx}(t,x)\right\}, \\ & \text{for all } \ t,x \in [0,T]\times I\!R^n. \\ v(T,x) = & \ \ell_F(x), \ \text{for all } \ x \in I\!R^n. \end{aligned} \tag{HJB_p}$$

where by $a_p t, x, u)$ we denote the projection of $at, x, u)$ on the cone $\mathcal{C}(\mathcal{S}_p)$. Denote by v_p the (well-defined) corresponding solution. When the step sizes vanish the limit of error between the solution of HJB and the one of the scheme is $\|v - v_p\|_\infty$. Using [8] we can obtain estimates of $\|v - v_p\|_\infty$. For infinite horizon problems we can obtain similar results applying [1, lemma 2.6].

7 Numerical results

We have implemented the algorithm in the C programming language and tested it on two academic examples in which the value function is known. Also we integrate on a finite rectangular domain with exact values on the boundary. This allows to compute the error made by the scheme and to see if its behavior is in agreement with the theory. For points of the grid close to the boundary, the size of the stencil may be smaller than p_{max} since points out of the domain are not used. Therefore, in the vicinity of the boundary the errors of approximation of covariances are larger than far from the boundary.

We use the reverse-time function $W(s, x) = V(T - s, x)$ in order to integrate t from 0 to T.

7.1 An uncontrolled problem

Our first test function is

$$\begin{cases} W(t, x_1, x_2) = (1 + t) \sin x_1 \sin x_2 \\ 0 \leqslant x_1 \leqslant \pi; \quad 0 \leqslant x_2 \leqslant \pi; \quad 0 \leqslant t \leqslant 1. \end{cases} \tag{7.17}$$

We choose $\Delta x := h_1 = h_2$, $N_1 h_1 = N_2 h_2 = \pi$, and the measurement of error is the mean value in L^1 norm, i.e. $e := \dfrac{\|W_{approx} - W_{exact}\|_1}{N_1 \times N_2}$. The following expressions for ℓ, f and σ are compatible with the HJB equation:

$$\begin{cases} \ell(t, x_1, x_2) = & \sin x_1 \sin x_2 [1 + (1 + 2\beta)(1 + t)] \\ & -2(1 + t) \ \cos x_1 \cos x_2 \sin(x_1 + x_2) \cos(x_1 + x_2) \\ f(t, x_1, x_2) = & 0 \\ a(t, x_1, x_2) = & \begin{pmatrix} \sin^2(x_1 + x_2) + \beta^2 & \sin(x_1 + x_2) \cos(x_1 + x_2) \\ \sin(x_1 + x_2) \cos(x_1 + x_2) & \cos^2(x_1 + x_2) + \beta^2 \end{pmatrix} \end{cases}$$

here $\sigma(t, x_1, x_2) = \begin{pmatrix} \sin(x_1 + x_2) & \beta & 0 \\ \cos(x_1 + x_2) & 0 & \beta \end{pmatrix}$.

We display in figure 8 the logarithm of error function of discretization step, for $\beta^2 = 0.1$ and 0, when $p_{max} = 5$. The scheme is consistent only in the first case. Accordingly, the error decreases when the space step is reduced in the first case, but not in the other.

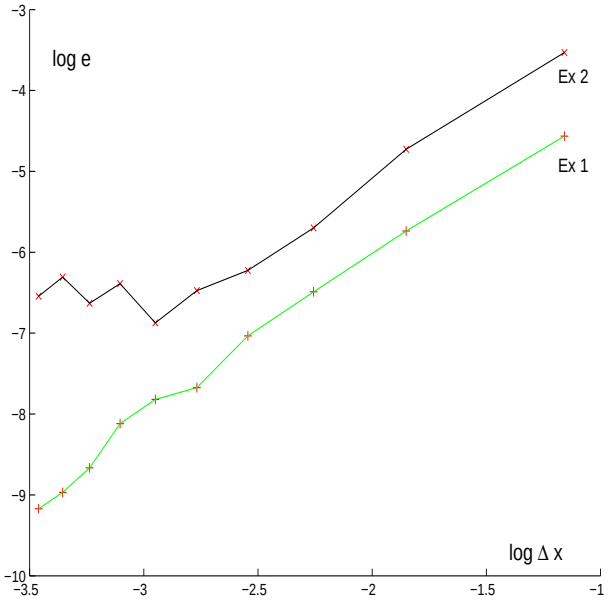

Figure 8: Error vs discretization step, $p_{max} = 5$

7.2 Numerical example, optimal control

We consider here an optimal control problem where $\sigma(\cdot)$ and $a(\cdot)$ do not depend on the control. Also, the trend is $f(t, x, u) = u$, with restriction $u_1^2 + u_2^2 \leq 1$.
The test function is

$$\left\{ \begin{array}{l} W(t, x_1, x_2) = (1 + t) \sin x_1 \sin x_2 \\ -1 \leqslant x_1 \leqslant 1; \quad -1 \leqslant x_2 \leqslant 1; \quad 0 \leqslant t \leqslant 0.5 \end{array} \right. \tag{7.18}$$

We have here a degenerate diffusion $a(t, x_1, x_2) = \frac{1}{2}\sigma(t, x_1, x_2)\sigma(t, x_1, x_2)^\top$ with

$$\sigma_1(t, x_1, x_2) = \sqrt{2}\ \sin(x_1 + x_2),\ \ \sigma_2(t, x_1, x_2) = \sqrt{2}\ \cos(x_1 + x_2)$$

The resulting distributed cost is

$$\begin{aligned}\ell(t,x_1,x_2) = & \sin(x_1)\sin(x_2) \\ & +(1+t)\Big[\big(\cos^2(x_1)\sin^2(x_2)+\sin^2(x_1)\cos^2(x_2)\big)^{1/2} \\ & +\sin(x_1)\sin(x_2) \\ & -2\sin(x_1+x_2)\cos(x_1+x_2)\cos(x_1)\cos(x_2)\Big]\end{aligned}$$

We display in figure 9 the error, as defined in section 7.1, vs the discretization step when $p_{max} = 10$. Although the scheme is not consistent, it appears that the discretization errors are quite small.

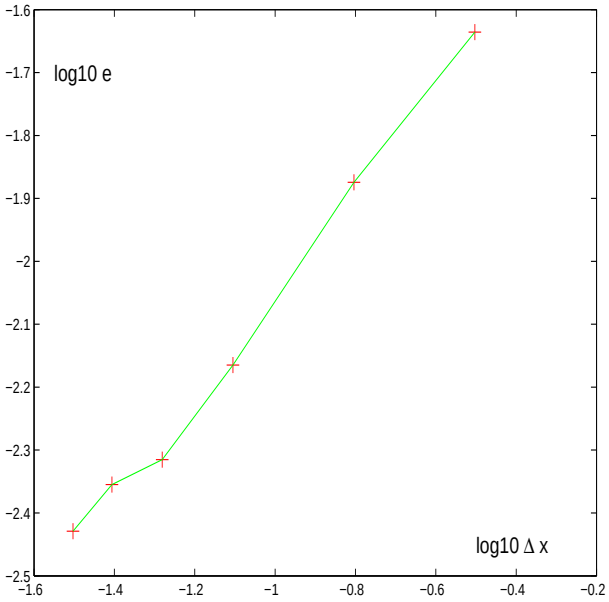

Figure 9: Error vs discretization step, optimal control, $p_{max} = 10$

References

[1] G. Barles and E.R. Jakobsen. On the convergence rate of approximation schemes for Hamilton-Jacobi-Bellman equations. *M2AN. Mathematical Modelling and Numerical Analysis*, 36:33–54, 2002.

[2] G. Barles and E.R. Jakobsen. Error bounds for monotone approximation schemes for Hamilton-Jacobi-Bellman equations. To appear, 2003.

[3] G. Barles and P. E. Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic Analysis*, 4:271–283, 1991.

[4] J. F. Bonnans and H. Zidani. Consistency of generalized finite difference schemes for the stochastic HJB equation. *SIAM J. Numerical Analysis*, 41:1008–1021, 2003.

[5] F. Camilli and M. Falcone. An approximation scheme for the optimal control of diffusion processes. *RAIRO Modélisation Mathématique et Analyse Numérique*, 29:97–122, 1995.

[6] W.H. Fleming and H.M. Soner. *Controlled Markov processes and viscosity solutions*. Springer, New York, 1992.

[7] R.L. Graham, D.E. Knuth, and O. Patashnik. *Concrete Mathematics, A Foundation For Computer Science*. Addison-Wesley, Paris, 1994.

[8] E.R. Jakobsen and K.H. Karlsen. Continuous dependence estimates for viscosity solutions of fully nonlinear degenerate parabolic equations. To appear, 2003.

[9] N.V. Krylov. On the rate of convergence of finite-difference approximations for Bellman's equations with variable coefficients. *Probability Theory and Related Fields*, 117(1):1–16, 2000.

[10] H.J. Kushner. *Probability methods for approximations in stochastic control and for elliptic equations*. Academic Press, New York, 1977. Mathematics in Science and Engineering, Vol. 129.

[11] H.J. Kushner and P.G. Dupuis. *Numerical methods for stochastic control problems in continuous time*, volume 24 of *Applications of mathematics*. Springer, New York, 2001. Second edition.

[12] P.-L. Lions. Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations. Part 2: viscosity solutions and uniqueness. *Communications in partial differential equations*, 8:1220–1276, 1983.

[13] P.-L. Lions and B. Mercier. Approximation numérique des équations de Hamilton-Jacobi-Bellman. *RAIRO Analyse numérique*, 14:369–393, 1980.

[14] J.-L. Menaldi. Some estimates for finite difference approximations. *SIAM J. Control Optim.*, 27:579–607, 1989.

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399